文章编号：1006-3080(2025)03-0363-08 DOI：10.14135/j.cnki.1006-3080.20240731003

基于多层级领域自适应网络和置信度约束的道路场景语义分割方法

万才路，堵　威
（华东理工大学能源化工过程智能制造教育部重点实验室，上海 200237）

摘要：语义分割旨在为图像中的每个像素分配一个类别标签，在自动驾驶领域具有广泛的应用。在实际应用中，在某个场景训练的语义分割模型无法有效地应用到其他场景，这成为实际应用中的一个关键问题，像素级域内自适应方法已被证明是一种解决此问题的有效方法。然而，这种方法不能有效地利用空间位置信息，并且容易受到噪声伪标签的影响。为了解决这些问题，本文首先提出了多层级领域自适应网络和置信度约束的方法，同时利用空间先验知识提出了阈值方法，提高了在域内自适应中使用伪标签的质量。在“GTA5 到 Cityscapes”和“SYNTHIA 到 Cityscapes”任务中，相较于基准方法本文所提方法分别实现了 6.5% 和 2.8% 的性能提升。

关键词：道路场景；语义分割；领域自适应；自训练；对抗学习

中图分类号：TP391.4　　**文献标志码：**A

近年来研究热点的自动驾驶要求对复杂城市街道拥有强大的场景理解能力，以做出决策并控制运动系统[1]。协助车辆识别其周围环境的常见方法是对车辆上安装的摄像头拍摄图像的应用语义分割。语义分割将图像中的所有像素分类为类别标签，是计算机视觉中的重要任务之一，许多方法[2-6]被提出来改进性能并取得了惊人的效果。然而，这些方法的关键限制是它们需要大量高质量的标签，这需要大量的人力和物质资源。例如，来自 Cityscapes[7] 数据集的一张图像需要一个人花费 90 min 来标注。从虚拟图像引擎中收集的合成数据集[8-9]被用来应对这个限制，因为这些图像引擎可以自动导出图像和对应标签。然而，由于现实世界场景和虚拟世界场景之间的外观差异很大，因此尽管模型在虚拟世界中做出高精度预测，但将其应用于现实世界图像的类别标签预测仍然困难。

解决有标签的虚拟世界图像（源域）和没有标签的现实世界图像（目标域）之间的领域差异问题的无监督领域自适应是关键技术之一。近来无监督领域自适应通过对抗学习或者自训练的方法来减小不同领域之间数据分布的差异[10-14]。对抗学习通过欺骗领域判别器来实现源域和目标域分布的全局对齐。自训练则是循环迭代目标域图像的伪标签并采用置信度估计、一致性正则化或熵最小化等方法来提高分割性能，伪标签即高可信度的预测。Pan 等[15]提出领域差异不仅存在于源域和目标域之间（称为域间差异），还存在于目标域的不同部分之间（称为域内差异），研究使用基于熵的排序方法将目标图像分为“容易”或“困难”两部分，并实现了图像级域内自适应。Yan 等[16] 认为仅在图像级别进行域内自适应是不够的，因为语义分割网络分配的是像素级别的类别标签。因此，研究提出了一个两步的无监督领域自适应方法，以实现像素级域内自适应。具体来说，在训练一个域间自适应网络 AdaptSegNet [12]后，通过一种类别阈值方法将目标域图像的像素分为“容易”或“困难”两部分。该阈值方法基于来自 AdaptSegNet[12] 预测的置信度分数为每个类别选择阈值，高出这个阈值的像素为“容易”像素，低于这个阈值的像素为“困难”像素，利用“容易”像素的伪标签，将分割网络从“困难”像素适应到“容易”像素以提高对“困难”像素

收稿日期：2024-07-31

作者简介：万才路（2000—），男，安徽人，硕士生，主要研究方向：计算机视觉。E-mail：y30210979@mail.ecust.edu.cn

通信联系人：堵　威，E-mail：duwei0203@ecust.edu.cn

引用本文：万才路，堵　威. 基于多层级领域自适应网络和置信度约束的道路场景语义分割方法 [J]. 华东理工大学学报（自然科学版），2025，51(3)：363-370.

Citation: WAN Cailu, DU Wei. Semantic Segmentation Methods for Road Scenes Based on Multi-Level Domain Adaptation Network and Confidence Constraints[J]. Journal of East China University of Science and Technology, 2025, 51(3): 363-370.

的预测精度。然而,像素级域内自适应无法有效地利用图像的空间位置信息并且对伪标签的质量非常敏感。因为伪标签在捕捉空间布局方面存在不足,这导致分割网络忽略了关键的空间位置信息。此外,像素级域内自适应过程对伪标签的依赖性过高,这意味着带有噪声的伪标签可能会严重阻碍网络对某些类别的有效学习。

为了解决上述问题,本文提出了3种方法:首先,提出了一个多层级领域自适应网络,旨在同时减少图像级别和像素级别的分布差异。鉴于图像中丰富的空间位置信息,本文在像素级领域内自适应的基础上引入了图像级领域内自适应。通过图像级对抗学习策略,促使目标域和源域在空间布局上的预测趋于一致,从而显著提升了预测的准确性;其次,提出了一种基于置信度约束的方法,以减轻伪标签对分割网络性能的负面影响。与之前方法不同,本文不仅将像素分类为"容易"或"困难",还记录了"容易"类别中像素的置信度值。通过引入置信度损失函数,有效地约束了网络在域内自适应过程中对伪标签的过度拟合;最后,通过整合空间先验知识,改进了现有的类别阈值方法,以降低伪标签的错误率。这种方法利用了源域中类别频率的空间结构相似性,从而提高了伪标签的整体质量。

1 基于多层级领域自适应网络和置信度约束的道路场景语义分割方法

本文使用 S 表示源域,包含合成图像 X_s 和标签 Y_s,共有 C 个不同的类别;用 T 表示目标域,其包含真实图像 X_t。

1.1 像素分离

如图1所示,为了适应域内差异,首先将目标图像的像素分为"容易"和"困难"两部分。置信度值可以衡量来自语义分割模型的预测 P 的准确性,具有高置信度值的预测往往比置信度值较低的预测更准确。基于这一规则,许多方法中使用置信度值大于0.9的像素预测作为伪标签[17]。由于观察到源域和目标域的空间位置信息上存在显著的相似性,如图2(a)所示,源域和目标域的图像在视觉外观上可能截然不同,但它们在空间位置信息上却存在显著的相似性。例如,天空通常位于图像的顶部,而汽车则总是出现在道路上。为了利用分割中的空间先验知识,本研究统计源域图像中每个类别在空间上的分布,如图2(b)所示。本文用 $F_s^{(h,w,c)}$ 表示源域图像中像素 (h,w) 处类别 c 的频率:

$$F_s^{(h,w,c)}=\frac{N_c^{(h,w)}}{\sum_{c=1}^{C}N_c^{(h,w)}} \tag{1}$$

其中 $N_c^{(h,w)}$ 是类别 c 出现在像素 (h,w) 处的次数,$\sum_{c=1}^{C}N_c^{(h,w)}$ 表示所有类别出现在像素 (h,w) 处的次数。本文使用一个 $n\times n$ 的高斯核对 $F_s^{(h,w,c)}$ 进行平滑处理,将预测的置信度分数乘以频率,并将结果作为像素分离的参考。为了缓解类别不平衡问题,本文对每个类别计算像素分离的阈值。具体来说,给定目标域图像 x_t,将其输入预训练模型以获得预测。然后使用二进制掩码 $M_{x_t}\in\{0,1\}^{H\times W}$ 展示了在 x_t 中分离像素的结果:

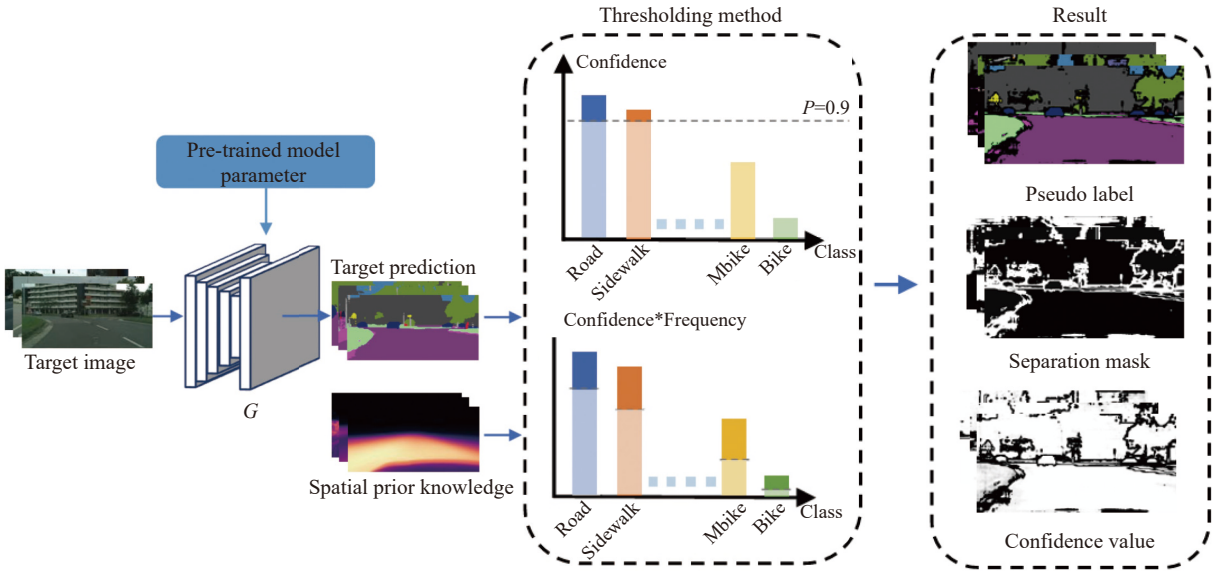


图1 像素分离

Fig. 1 Pixel separation

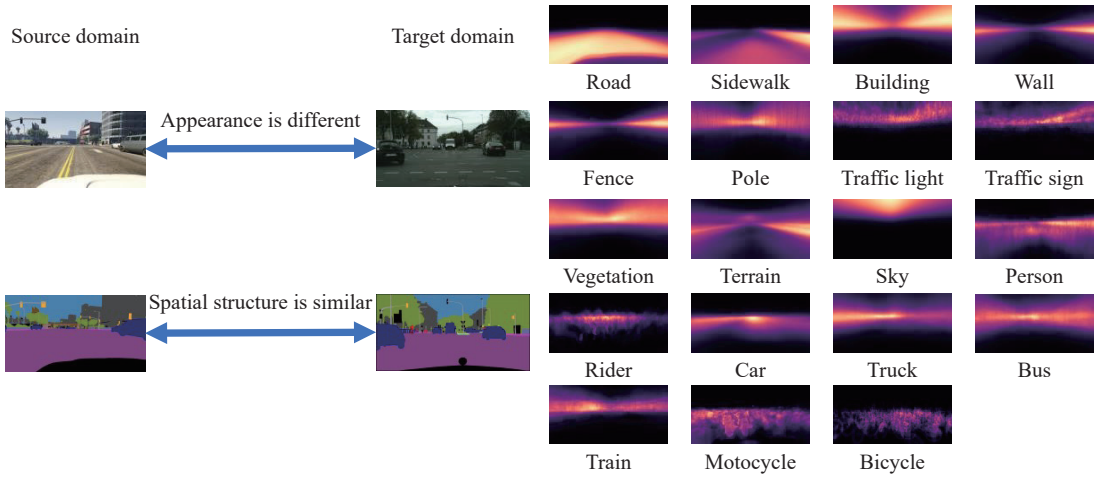


图 2　不同域图像的外观和空间结构 (a), 源域统计出的空间位置分布 (b)

Fig. 2　Appearance and spatial structure of images from different domains (a), Spatial position distribution derived from the source domain (b)

$$M_{x_t}^{(h,w)}=\begin{cases}0, & \text{if } \operatorname{argmax}_{\tilde{c}} P_{x_t}^{(h,w,\tilde{c})}=c \\ & \text{and } P_{x_t}^{(h,w,c)}F_s^{(h,w,c)}>t^{(c)} \\ 0, & \max P_{x_t}^{(h,w)}>0.9 \\ 1, & \text{otherwise}\end{cases} \tag{2}$$

其中, $t^{(c)}$ 是类别 c 的像素分离阈值。具体而言, $r^{(c)}$ 表示在整个数据集上将类别 c 的置信度分数乘以频率的所有结果。 $t^{(c)}$ 被确定为使得 $r^{(c)}$ 中大于 $t^{(c)}$ 的元素数量等于 $\alpha\cdot\left|r^{(c)}\right|$, 其中 α 表示 “容易”像素的比例, $\left|r^{(c)}\right|$ 表示 $r^{(c)}$ 的长度。 $P_{x_t}^{(h,w)}$ 表示 $x_t^{(h,w)}$ 的预测。 $M_{x_t}^{(h,w)}=0$ 表示像素为“容易”部分, $M_{x_t}^{(h,w)}=1$ 表示像素为“困难”部分。根据 $M_{x_t}^{(h,w)}$ 分配伪标签 $\hat{y}_{x_t}^{(h,w)}$, 同时记录置信度 $\hat{p}_{x_t}^{(h,w)}$, 如下:

$$\begin{aligned}\hat{y}_{x_t}^{(h,w)}&=\begin{cases}\operatorname{argmax}_{\tilde{c}} P_{x_t}^{(h,w,\tilde{c})}, & M_{x_t}^{(h,w)}=0\\ \text{none}, & M_{x_t}^{(h,w)}=1\end{cases}\\ \hat{p}_{x_t}^{(h,w)}&=\begin{cases}\max P_{x_t}^{(h,w)}, & M_{x_t}^{(h,w)}=0\\ \text{none}, & M_{x_t}^{(h,w)}=1\end{cases}\end{aligned} \tag{3}$$

值得注意的是, 为了获得伪标签和置信度值, 需要一个预训练模型。为了更好地与 PixIntraDA[16] 进行比较, 本文使用了与 PixIntraDA[16] 相同的预训练模型来生成伪标签。

1.2　多层级领域自适应网络和置信度约束

为了同时减小域间差异和域内差异, 以及避免过度拟合带有噪声的伪标签, 本文方法包括多层级领域自适应网络和置信度约束。

1.2.1　*多层级领域自适应网络*　本文的多层级领域自适应网络包含像素级域内自适应以及图像级域间自适应。如图 3 所示, 在源域上, 网络 G 接受一个带有标签 y_{x_s} 的图像 x_s 作为输入, 并生成预测 $P_{x_s}=G(x_s)$。在目标域上, 由于图像缺乏真实标签, 本文使用通过式 (3) 选定的伪标签 $\hat{y}_{x_t}$ 来监督网络训练。具体来说, 网络 G 接受一个带有伪标签 $\hat{y}_{x_t}$ 的图像 x_t 作为输入, 并生成预测 $P_{x_t}=G(x_t)$。通过最小化交叉熵损失 $\mathcal{L}_{\text{seg}}$ 来优化 G:

$$\begin{aligned}\mathcal{L}_{\text{seg}}=&-\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} y_{x_s}^{(h,w,c)}\lg\left(P_{x_s}^{(h,w,c)}\right)-\\ &\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}\hat{y}_{x_t}^{(h,w,c)}\lg\left(P_{x_t}^{(h,w,c)}\right)\end{aligned} \tag{4}$$

公式 (3) 中赋予“none”值伪标签 $\hat{y}_{x_t}$ 表明相应的像素不参与分割损失的计算。由于预训练模型效果有限, 它不可避免地产生许多错误预测, 这些错误预测不能用作伪标签。为了避免这些错误预测的不良影响, 本文为它们分配“none”值。

为了学习源域图像中类别的空间布局, 本文利用对抗学习来使 P_{x_t} 和 P_{x_s} 的分布对齐。具体来说, D_{inter} 被训练区分来自源域图像或目标域图像的预测, 而 G 被训练为源域图像和目标域图像生成的相似预测, 以欺骗 D_{inter}。因此, D_{inter} 和 G 的优化问题被表述如下:

$$\begin{aligned}\mathcal{L}_{\text{inter adv}}^{D_{\text{inter}}}&=-\lg\left(1-D_{\text{inter}}(P_{x_t})\right)-\lg\left(D_{\text{inter}}(P_{x_s})\right)\\ \mathcal{L}_{\text{inter adv}}^{G}&=-\lg\left(D_{\text{inter}}(P_{x_t})\right)\end{aligned} \tag{5}$$

此外, 为了解决“容易”和“困难”像素之间的域内差异, 按照 PixIntraDA[16] 中提出的像素级对抗学习方法, 域内判别器 D_{intra} 被训练为区分像素是来自“容易”还是“困难”部分, 而 G 被训练为欺骗 D_{intra}。因此, 用于优化 D_{intra} 和 G 的像素级对抗损失被表述为

$$\begin{aligned}\mathcal{L}_{\text{intra adv}}^{D_{\text{intra}}}=&-\lg\left[\mathbf{1}_H^T\left(D_{\text{intra}}(G(x_t))\odot M_{x_t}\right)\mathbf{1}_W\right]-\\ &\lg\left[\mathbf{1}_H^T\left((\boldsymbol{J}-D_{\text{intra}}(G(x_t)))\odot(\boldsymbol{J}-M_{x_t})\right)\mathbf{1}_W\right]\\ \mathcal{L}_{\text{intra adv}}^{G}=&-\lg\left[\mathbf{1}_H^T\left((\boldsymbol{J}-D_{\text{intra}}(G(x_t)))\odot M_{x_t}\right)\mathbf{1}_W\right]\end{aligned} \tag{6}$$

其中, $\mathbf{1}_H$ 和 $\mathbf{1}_W$ 分别表示大小为 $H\times1$ 和 $W\times1$ 的全

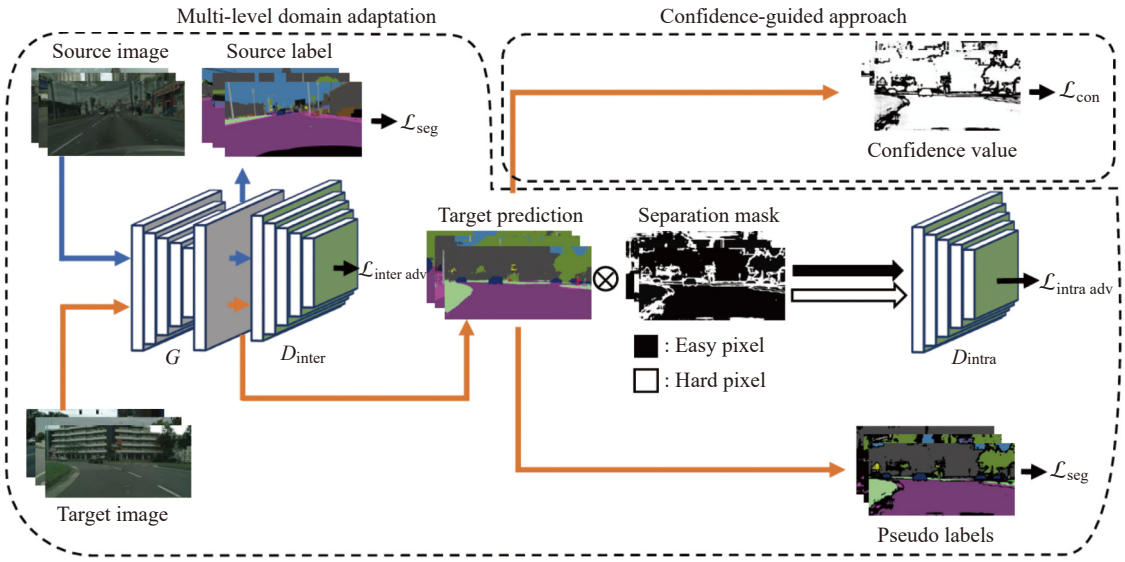


图3 多层级领域自适应网络和置信度约束

Fig. 3 Multi-level domain adaptive network and confidence regularization

一向量，$\boldsymbol{J}$ 表示大小为 $H\times W$ 的全一矩阵，$\odot$ 表示 Hadamard 积。本文的多层级领域自适应与像素级域内自适应 PixIntraDA[16] 的主要区别：本文通过对抗学习引导网络学习源域的空间布局和目标域“容易”像素的置信度分布，而像素级域内自适应仅学习“容易”像素的置信度分布。

1.2.2 置信度约束 为了避免过拟合带有噪声的伪标签，本文借助置信度 $\widehat{p}_{x_t}$、伪标签 $\widehat{y}_{x_t}$ 以及目标预测 P_{x_t} 引入置信度损失 $\mathcal{L}_{\text{con}}$。具体来说，置信度损失定义如下：

$$\mathcal{L}_{\text{con}}=\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}\hat{y}_{x_t}^{(h,w,c)}\left|\hat{p}_{x_t}^{(h,w,c)}-P_{x_t}^{(h,w,c)}\right| \tag{7}$$

其中，$\widehat{p}_{x_t}^{(h,w,c)}$ 是将 $\widehat{p}_{x_t}^{(h,w)}$ 转换为与 $\widehat{y}_{x_t}^{(h,w,c)}$ 相同大小的结果，对于所有 $\forall c\in C,\widehat{p}_{x_t}^{(h,w,c)}=\widehat{p}_{x_t}^{(h,w)}$。

因此，用于优化 G 的完整损失函数形式为：

$$L=\mathcal{L}_{\text{seg}}+\lambda_{\text{con}}\mathcal{L}_{\text{con}}+\lambda_{\text{inter}}\mathcal{L}_{\text{inter adv}}^{G}+\lambda_{\text{intra}}\mathcal{L}_{\text{intra adv}}^{G} \tag{8}$$

其中，λ_{con}、λ_{inter} 和 λ_{intra} 分别表示置信度损失 $\mathcal{L}_{\text{con}}$、域间对抗损失 $\mathcal{L}_{\text{inter adv}}^{G}$ 和域内对抗损失 $\mathcal{L}_{\text{intra adv}}^{G}$ 的权重。

2 实验与结果分析

2.1 实验部分

2.1.1 数据集 Cityscapes[7] 是从50个不同的城市收集的大规模数据集，包含5000张带有像素级语义标注的图像和20000张带有粗略语义标注的图像。本文使用该数据集中来自训练集的2975张无标签图像作为目标域，来自 Cityscapes 验证集的500张带有像素级语义标注的图像来评估训练的模型。GTA5[8] 包括24966个分辨率为1914×1052的图形，是由虚拟游戏引擎生成。GTA5 和 Cityscapes之间有19个相同类别，因此本文使用这些相同的类别并忽略其他类别来训练网络。SYNTHIA[9] 包含9400张带有高质量掩码的城市场景图片，是一个合成数据集。在训练时，SYNTHIA 有13个与 Cityscapes类兼容的类别。

2.1.2 评估 语义分割模型的性能通过平均交并比(mIoU)指标进行评估，这在语义分割的无监督领域自适应中经常用于与其他模型进行比较。

2.1.3 网络架构 本文采用了 Deeplab-V2[2] 作为语义分割的架构，其架构骨干是在 ImageNet[18] 上预训练过的 ResNet-101[19]，并在训练中进行参数微调；域间自适应对从第5层卷积的输出特征执行自适应。与语义分割架构相对应，加入鉴别器(与 DCGAN[20] 中使用的架构相同)执行对抗学习，以对齐来自第5层卷积预测的空间分布。在域内自适应中，为了实现像素级对抗学习，使用与 PixIntraDA[16] 中相同的鉴别器，并对鉴别器生成的输出进行双线性采样，使其与输入图像的大小相同。

2.1.4 算法的实现细节 本文所有实验中使用 PyTorch 深度学习框架，实验环境为一块搭载有24 GB内存的 NVIDIA GeForce RTX 3090 GPU。在多层级领域自适应中，收集将源域图像转换为目标域风格的图像。同时，本文使用一个从 PixIntraDA[16] 的域间自适应中训练得到的预训练模型产生伪标签，批量大小为4，在框架中采用多尺度训练和测试。此外，λ_{intra} 设置为0.05，λ_{inter} 设置为0.01。为了实现主要适应领域内差异的目标，本文根据 $\frac{\lambda_{\text{intra}}}{\lambda_{\text{inter}}}=5$ 确定

λ_{inter}，与主流方法保持一致[11-12, 15-16]，λ_{con} 通过实验确定。

2.1.5　算法复杂度　算法的时间、空间复杂度分别如公式(9)和公式(10)所示：

$$\text{Time} \sim O\left(\sum_{l=1}^{D} M_l^2 \cdot K_l^2 \cdot C_{l-1} \cdot C_l\right) \tag{9}$$

$$\text{Space} \sim O\left(\sum_{l=1}^{D} K_l^2 \cdot C_{l-1} \cdot C_l + \sum_{l=1}^{D} M_l^2 \cdot C_l\right) \tag{10}$$

其中，M_l，K_l 和 C_l 分别代表第 l 层网络的输出特征图大小、卷积核大小和输出通道数，D 代表网络卷积层的数量。具体来说，本文算法模型的浮点运算量约为 3.74×10^{11} 次，模型参数的大小约为 170 MB。此外，模型训练在使用单块 NVIDIA GeForce RTX 3090GPU 的情况下需要耗时 20 h。

2.2　结果分析

2.2.1　定量结果　表 1 和表 2 所示为本文的方法与其他方法的比较结果。为了确保公平性和一致性，除了 TTA(Test Time Argument action)外，所有方法均采用了 Deeplab-V2 作为语义分割的基础架构。TTA 采用扩散模型作为其语义分割架构，该模型的网络规模超过了 Deeplab-V2。总体而言，本文的方法在任务“GTA5 到 Cityscapes”和“SYNTHIA 到 Cityscapes”中 mIoU 分别提高到 52.6% 和 56.0%。与表 1 和表 2 中的基线 PLA 相比，本文的方法在任务“GTA5 到 Cityscapes”和“SYNTHIA到 Cityscapes”中 mIoU 分别提高了 6.5% 和 2.8%。本文方法在有效性上与其他最先进的方法也具有相当的竞争力。由于 Cityscapes 和 SYNTHIA 之间在空间布局上存在相对较大的差异，在任务“SYNTHIA 到 Cityscapes”中没有使用空间先验知识。需要注意的是，PixIntraDA[16] 包含像素级对抗学习、连续索引对抗学习、多一轮训练和 Kullback-Leibler 正则化 4 个部分。由于本文的工作基于像素级对抗学习部分，而不包含其他 3 个部分，为了更好地突显本文方法的有效性，主要关注了本文方法与像素级对抗学习部分的比较。PixIntraDA[16] 方法中像素级对抗学习多尺度测试结果 mIoU 是 49.4%。本文通过图 4 所示的分割结果可视化来比较本文方法与像素级对抗学习部分的有效性。

2.2.2　消融实验　本文的方法包含 3 个部分，多层级领域自适应(Multi-Level Domain Adaptation, MDA)、置信度正则化(Confidence Regularization, CR)和融入空间先验的阈值方法(Threshold Method Incorporating Spatial Prior, TMISP)。与 PLA 相比，本文方法在 3 个方面有所不同：首先，基于 PLA，本文提出了 MDA 来解决 PLA 不能有效利用图像的空间位置信息的问题；其次，CR 解决 PLA 可能过拟合噪声伪标签的问题；最后，新的阈值方法来获得更好的伪标签。表 3 验证了所有部分的有效性，可以观察到本文方法的所有部分都提高了有效性。

2.2.3　超参数分析　本文通过实验来选择 λ_{con}、α 和 n 的最优值，如表 4、表 5 和表 6 所示。为了确定最优的超参数组合，本文中采用了逐步超参数优化策略。初始 λ_{con}、α 和 n 值分别为 10、0.67 和 69，通过依次固定两个超参数，并优化剩余的一个超参数，使得优化过程更加直接和容易管理。在每一步的优化中，记录了不同配置下的性能指标，并基于这

表 1　GTA5 到 Cityscapes 数据集的分割结果

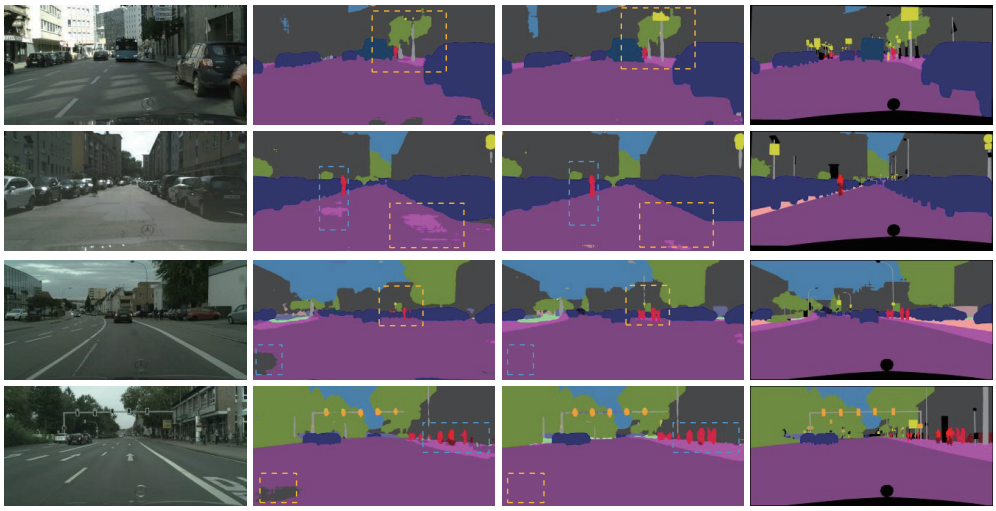
Table 1　Segmentation results of GTA5 to Cityscapes dataset

Method	mIoU/%																			
	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetable	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Mbike	Bike	Average
AdaptSetNet[12]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.2
AdvEnt[11]	89.9	36.5	81.6	29.2	25.2	28.5	32.3	22.4	83.9	34.0	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8
IntraDA[15]	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0	30.2	35.8	46.3
ASDM[21]	89.0	36.2	84.1	42.7	25.4	29.4	35.4	16.7	84.4	29.6	80.0	58.1	**35.7**	85.0	34.0	36.5	**20.6**	29.3	39.6	46.9
BDL[17]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
PLA[16]	91.9	53.1	84.8	30.9	26.6	35.5	41.0	28.5	85.7	38.7	87.1	66.0	28.6	83.9	31.9	35.3	0.1	37.4	**50.7**	49.4
LRTIR[22]	**92.9**	**55.0**	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
FDA[23]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	**38.9**	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
TTA[24]	86.1	36.2	**87.8**	**47.0**	**39.9**	**42.0**	46.1	36.5	**87.9**	**49.5**	**90.1**	64.6	28.4	**88.2**	39.4	**54.2**	3.3	31.3	32.7	52.2
Ours	91.6	50.6	85.6	36.7	30.6	38.3	**47.2**	33.4	86.3	41.6	86.1	**68.2**	34.0	86.1	**42.7**	51.4	1.3	**38.0**	48.8	**52.6**

表 2 SYNTHIA 到 Cityscapes 数据集的分割结果

Table 2 Segmentation results of SYNTHIA to Cityscapes dataset

Method	mIoU/%													
	Road	Sidewalk	Building	Light	Sign	Vegetable	Sky	Person	Rider	Car	Bus	Mbike	Bike	Average
AdaptsegNet[12]	84.3	42.7	77.5	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	46.7
AdvEnt[11]	85.6	42.2	79.7	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	48.0
IntraDA[15]	84.3	37.7	79.5	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	48.9
LRTIR[22]	**92.6**	**53.2**	79.2	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	49.3
BDL[17]	86.0	46.7	80.3	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	51.4
FDA[23]	79.3	35.0	73.2	**19.9**	24.0	61.7	82.6	61.4	31.1	83.9	40.8	38.4	51.1	52.5
PLA[16]	88.9	52.8	81.0	7.2	13.8	80.8	84.3	61.6	35.3	74.3	42.4	39.6	46.9	54.5
TTA[24]	87.2	46.6	**86.0**	18.3	**34.0**	**85.0**	**90.0**	61.7	28.9	**85.7**	**44.3**	26.3	26.5	55.4
Ours	83.9	47.6	77.0	15.4	16.6	81.7	85.7	**68.3**	**35.8**	80.6	42.4	**45.3**	**47.8**	**56.0**



(a) Input (b) PLA (c) Ours (d) Ground truth

图 4 分割结果

Fig. 4 Segmentation results

表 3 消融实验

Table 3 Ablation experiment

Method	TMISP	CR	MDA	mIoU/%
Full method	√	√	√	52.6
W/O TMISP		√	√	52.2
W/O TMISP+CR			√	51.4
W/O TMISP+CR+MDA(PLA)				49.4

表 4 λ_{con} 参数分析

Table 4 Analysis of λ_{con}

λ_{con}	mIoU/%
0	51.4
1	51.7
8	51.8
9	52.2
10	52.1

些数据选择了表现最优的超参数配置。通过这种逐步调整的方法，最终确定的超参数设置在每一步中均表现最优，确保超参数选择的结果是 3 个参数的综合最优配置。结果表明，当 λ_{con} 、α 和 n 分别等于 9、0.5 和 69 时，可以获得最佳结果。此外，α 等于 1 意味着所有预测都被用作伪标签。因此，这表明对于不可靠的预测赋予“none”值，使相应的像素不参与计算分割损失是必要。

表 5　α 参数分析

Table 5　Analysis of α

α	mIoU/%
0.33	52.1
0.5	52.6
0.67	52.2
1	46.0

表 6　n 参数分析

Table 6　Analysis of n

n	mIoU/%
39	52.0
69	52.6
99	52.3

3　结束语

本文提出了一个多层级领域自适应网络以缩小域间差异和域内差异; 利用置信度约束的方法以缓解噪声伪标签的影响。为了进一步提高伪标签的质量, 本文将空间先验知识与现有的阈值方法结合起来以选择伪标签。实验结果表明本文方法的性能卓越, 这 3 种方法并非孤立存在, 而是相互补充和增强的。置信度约束方法通过记录置信度值来减小潜在错误伪标签的影响, 而改进的阈值方法则利用空间先验知识来提升伪标签的准确性。在未来的工作中, 有一些方法可以进一步提高分割效果。由于从模型生成的高质量伪标签有助于训练域自适应网络, 因此使用更有效模型来获取更好的伪标签可以进一步提高分割结果。此外, 由于选择用于域内自适应像素的分离方法也影响伪标签的质量, 因此找到更好的分离方法提高效果。不同形式的置信度损失可能会产生不同的效果, 因此可以进一步研究最佳形式。

参考文献:

[1] YURTSEVER E, LAMBERT J, CARBALLO A, *et al*. A survey of autonomous driving: Common practices and emerging technologies[J]. IEEE Access, 2020, 8: 58443-58469.

[2] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al*. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[3] CHEN L C, PAPANDREOU G, SCHROFF F, *et al*. Rethinking atrous convolution for semantic image segmentation[J/OL]. (2017-06-07)[2017-08-25]. https://arXiv preprint arXiv: 170605587.

[4] 李钰, 袁晴龙, 徐少铭, 等. 基于感知注意力和轻量金字塔融合网络模型的室内场景语义分割方法 [J]. 华东理工大学学报 (自然科学版), 2023, 49(1): 116-127.

[5] 吴骏逸, 谷小婧, 顾幸生. 基于可见光/红外图像的夜间道路场景语义分割 [J]. 华东理工大学学报 (自然科学版), 2019, 45(2): 301-309.

[6] 夏源祥, 刘渝, 楚程钱, 等. 基于子空间多尺度特征融合的试卷语义分割 [J]. 华东理工大学学报 (自然科学版), 2023, 49(3): 429-438.

[7] CORDTS M, OMRAN M, RAMOS S, *et al*. The cityscapes dataset for semantic urban scene understanding [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 3213-3223.

[8] RICHTER S R, VINEET V, ROTH S, *et al*. Playing for data: Ground truth from computer games[C]//Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2016: 102-118.

[9] ROS G, SELLART L, MATERZYNSKA J, *et al*. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2016: 3234-3243.

[10] HOFFMAN J, WANG D, YU F, *et al*. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation [J/OL]. (2016-12-08)[2017-02-21]. https://arXiv preprint arXiv: 161202649.

[11] VU T H, JAIN H, BUCHER M, *et al*. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2019: 2517-2526.

[12] TSAI Y H, HUNG W C, SCHULTER S, *et al*. Learning to adapt structured output space for semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2018: 7472-7481.

[13] WANG Y, PENG J, ZHANG Z. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation [C]//Proceedings of the IEEE International Conference on Computer Vision. USA: IEEE, 2021: 9092-9101.

[14] TIAN Y, ZHU S. Partial domain adaptation on semantic segmentation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(6): 3798-3809.

[15] PAN F, SHIN I, RAMEAU F, *et al*. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2020: 3764-3773.

[16] YAN Z, YU X, QIN Y, *et al*. Pixel-level intra-domain adaptation for semantic segmentation [C]//Proceedings of the 29th ACM International Conference on Multimedia. NY, USA: IEEE, 2021: 404-413.

[17] LI Y, YUAN L, VASCONCELOS N. Bidirectional learning for domain adaptation of semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2019: 6936-6945.

[18] DENG J, DONG W, SOCHER R, *et al*. Imagenet: A large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2009: 248-255.

[19] HE K, ZHANG X, REN S, *et al*. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2016: 770-778.

[20] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. (2015-11-19)[2015-12-28]. https://arXiv preprint arXiv: 151106434.

[21] LUO X, CHEN W, LIANG Z, *et al*. Adversarial style discrepancy minimization for unsupervised domain adaptation[J]. Neural Networks, 2023, 157: 216-225.

[22] KIM M, BYUN H. Learning texture invariant representation for domain adaptation of semantic segmentation [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). USA: IEEE, 2020: 12972-12981.

[23] YANG Y, SOATTO S. Fda: Fourier domain adaptation for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2020: 4085-4095.

[24] GONG R, DANELLJAN M, SUN H, *et al*. Prompting diffusion representations for cross-domain semantic segmentation[EB/OL]. (2023-07-05)[2023-09-14]. https://arXiv preprint arXiv: 230702138.

Semantic Segmentation Methods for Road Scenes Based on Multi-Level Domain Adaptation Network and Confidence Constraints

WAN Cailu, DU Wei

(Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China)

Abstract: Semantic segmentation aims to assign a class label to each pixel in an image and has a wide range of applications. Semantic Segmentation needs large numbers of high-quality labels, which requires a lot of manpower and material resources. Furthermore, a semantic segmentation model trained on one domain cannot generalize well to other domains, which becomes a key problem in its practical applications. Unsupervised pixel-level intra-domain adaptation for semantic segmentation has been proven to be an effective method to address the problem. However, this method cannot effectively exploit spatial location information and is adversely affected by noisy pseudo-labels. In this work, we propose a confidence-guided multi-level domain adaptation approach to solve the problem. Specifically, we propose a multi-level domain adaptation framework to reduce the differences between pixels and spatial location information of images simultaneously. Moreover, to avoid that overfitting pseudo-labels may degrade the performance of the segmentation network, we construct a confidence loss function to constrain the network training. And we propose a method of selecting pseudo-labels and achieving better results in acquiring high-quality pseudo-labels than existing methods. We demonstrate the effectiveness of our approach through synthetic-to-real adaptation experiments. Compared with the unsupervised pixel-level intra-domain adaptation for semantic segmentation, our method leads to 6.5% and 2.8% relative improvements in mean intersection-over-union on the tasks “GTA5 to Cityscapes” and “SYNTHIA to Cityscapes”, respectively.

Key words: road scene; semantic segmentation; unsupervised domain adaptation; self-training; adversarial learning

（责任编辑：王晓丽）